# Behavioral Graph Analysis of Internet Applications

Kuai Xu, Feng Wang
Arizona State University
Email: {kuai.xu, fwang25}@asu.edu

***Abstract*—Recent years have witnessed rapid growth of innovative and disruptive Internet services such as video streaming and peer-to-peer applications. As network traffic of these applications continues to grow, it has become a challenging task to understand their communication patterns and traffic behavior of end hosts engaging in these applications. This paper presents a novel approach based on behavioral graph analysis to study social behavior of Internet application traffic based on bipartite graphs and one-mode projection graphs. Through a vector of graph properties including coefficient clustering that capture social behaviors of end hosts, we discover the inherent clustered groups of Internet applications that not only exhibit similar social behavior of end hosts but also have similar characteristics in the aggregated traffic. In addition, we demonstrate the usage of the proposed behavioral graph analysis in detecting emerging applications and anomalous traffic patterns towards Internet applications.**

## I. Introduction

Recent years have witnessed rapid growth of innovative and disruptive Internet services such as video streaming and peer-to-peer (P2P) applications. Understanding these new applications is very important and critical for traffic engineering, network monitoring and security management since these emerging applications are popular targets for attackers to discover and exploit vulnerabilities and weakness. As network traffic of these applications continues to grow, it has become a challenging task to understand communication patterns of these applications and traffic behavior of end hosts engaging in these applications.

Many prior work have focused on studying traffic characteristics and network behavior of certain important applications such as Web and P2P applications. Some other studies are devoted to classify network traffic into different applications based on packet payloads, application ports or statistical patterns. However, few attempt has been made to systematically analyze and explore the similarity among Internet applications. Towards this end, this paper proposes a novel approach based on behavioral graph analysis to characterize and profile social "community" behaviors of end hosts engaging in the same Internet application.

In this paper, we use use bipartite graphs and one-mode projection graphs to model Internet backbone traffic and capture the social behavior of source and destinations IP addresses (end hosts) of Internet applications. Through coefficient clustering and other graph properties that measure the "small-world" properties of end hosts (source or destination) for given applications, we find that many Internet applications share non-trivial similarity on social behaviors of end hosts in the same dimensions, which leads us to intuitively apply clustering algorithms to group applications with similar graph properties.

Using a vector of graph properties including coefficient clustering that captures social behaviors of end hosts engaging in the same Internet applications, we discover the inherent clustered groups of Internet applications that not only exhibit similar social behaviors of end hosts but also have similar characteristics in the aggregated traffic. For each application cluster, we examine characteristics of the aggregated traffic such as host symmetry, fan-out degree of source IP addresses, the fan-in degree of destination IP addresses. The experiment results based on real Internet backbone traffic demonstrate similar traffic characteristics of Internet applications in the same clusters. This observation confirms that the clustering step based on coefficient clustering indeed identifies Internet applications with similar graph properties in the one-mode projection graphs that are generated from the underlying network traffic.

Finally, we demonstrate the usage of the proposed behavioral graph analysis in detecting emerging applications and anomalous traffic patterns towards Internet applications through case studies. For example, we find that a service application SSH (TCP PORT 22) falls in the same clusters with ports (TCP PORT 135 and TCP PORT 1433) that are associated with well-known vulnerabilities during one time window. The in-depth analysis reveals that during this specific time window, there exist several attackers who are performing certain scanning activities towards a number of random destination hosts on TCP port 22, which dramatically changes the graph properties of this port.

The contributions of this paper include

- We introduce bipartite graphs and one-mode projection graphs to capture and distinguish social behavior of end hosts engaging in the same Internet applications;
- We apply clustering algorithms to group Internet applications into distinctive clusters based on coefficient clustering and other graph properties of one-mode projection graphs build from the underlying network traffic;
- We demonstrate the usage of the proposed behavioral graph analysis in detecting anomalous traffic behavior and emerging applications.

The remainder of this paper is organized as follows. Section II discusses related work and the difference between this study and these prior work. Section III presents the proposed method based on bipartite graphs and one-mode projection graphs, as well as the clustering algorithm employed for discovering clusters of Internet applications. Section IV describes

the experiment results based on real Internet backbone traffic collected from a Tier-1 Internet service provider. Section V concludes this paper and outlines our future work.

## II. Related Work

Many recent research have been focused on individual Internet applications, such as Web [1], DNS [2], and HTTPS [3]. For example, [1] studies the co-locations of Web servers on the Internet and their corresponding authoritative DNS servers to discover the relationships among Web servers that form the Web. In [2], Schatzmann et al. develop an approach of identifying HTTPS mail traffic from netflow data, while [3] analyzes DNS query traffic for in-depth understanding of the overall network traffic and detects unusual or unwanted network traffic on edge networks. However, little attempt has been made to systematically study all Internet applications. Although [4] develops an application-aware traffic measurement and analysis system, its major interest is on accurate usage-based accounting. Different from these work, this study focuses on social behavior of Internet applications with an ultimate goal of understanding traffic behavior of Internet applications.

Network traffic behavior has been extensively studied in recent years [5], [6], [7], [8]. In [5] Wei at el. apply agglomerative algorithms to cluster end hosts into clusters based on host profiles that consist of a number of traffic features including daily destination number, daily byte number, average TTL, TCP and UDP ports, and aggregated statistics for each destination. In [6], Karagiannis et al. classify network traffic based on patterns of host behavior at the function, social and application levels. For the application level, [6] uses four major features (source address, destination address, source port, and destination port) in data traffic and additional flow characteristics (transport protocol and packet size) to build graphlets to capture and classify the behavior of Internet hosts, while [7] builds behavior profiles of end hosts using their traffic communication patterns. [8] generates traffic profiles of network prefixes through behavior analysis of aggregated network traffic at the BGP-prefix level.

Several studies apply graph-based analysis to examine Internet traffic [9], [10], [11], [12]. An early work [9] generates communication graphs for E-mail traffic and applies interest-clustering algorithms for uncovering groups of E-mail users that share similar interests or expertise. [10] introduces traffic dispersion graphs to capture the social-behavior of end hosts using the edges representing traffic interactions between source and destination hosts. In [11], Yu et al. propose traffic activity graphs to represent the social interactions between source and destination hosts engaging in specific types of applications. Different from these works, this study uses one-projection graphs to capture the social-behavior similarity among all source or destination hosts involving in the same applications. Using one-mode projection graphs to capture social behavior of end hosts is first introduced in our earlier work [12], which constructs bipartite graphs from host communication patterns and then to builds one-mode projection graphs to discover social-behavior similarity among end hosts in the same network prefixes. Different from [12], this paper explores one-mode projection graphs of source or destination hosts engaging in the same applications for discovering behavior similarities among Internet applications.

## III. Methodology

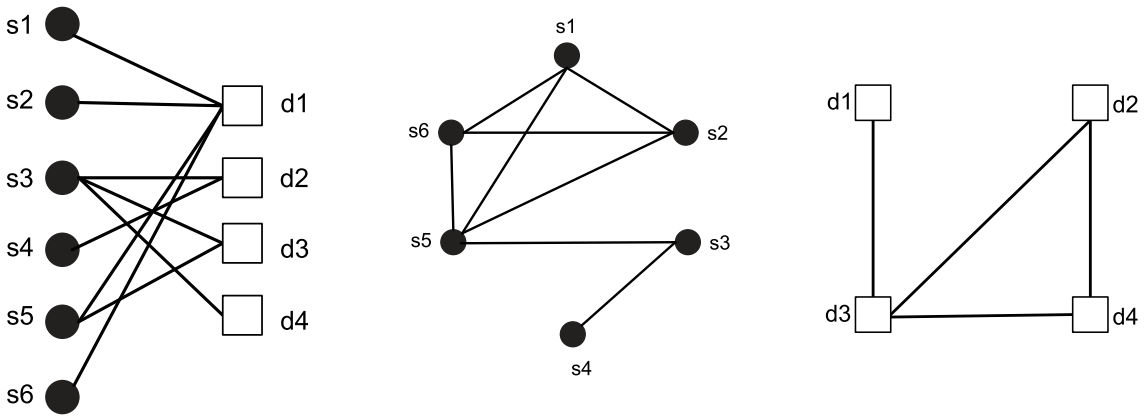
In this section we first describe how to use bipartite graphs and one-mode projection graphs to model Internet application traffic from backbone links. Subsequently, we adopt widely-used coefficient clustering and other graph properties of one-mode projection graphs to capture social behavior of end hosts engaging in the same applications. The clustered patterns of coefficient clustering lead us to apply a simple clustering step to group Internet applications with similar social-behavior into the same clusters.

### A. Modeling Application Traffic using Bipartite Graphs and One-Mode Projection Graphs

Data communications observed on Internet backbone links can naturally be represented by bipartite graphs where all IP packets originate from one set of nodes (i.e., source IP addresses) to another disjoint set of nodes (i.e., destination IP addresses). Let $G$ represent the bipartite graph to model such data communications, and let $\mathcal{S}$ and $\mathcal{D}$ represent the sets of source and destination IP addresses in the graph, respectively. For any given application port $p$, its traffic forms a subgraph of the bipartite graph, which could be represented with $G_p$. Similarly, we could use $\mathcal{S}_p$ and $\mathcal{D}_p$ to denote the sets of source and destination IP addresses engaging in the application port $p$.

Given a bipartite graph $G = (\mathcal{S}, \mathcal{D}, E)$, we could obtain one-mode projection graphs $G_{\mathcal{S}} = (\mathcal{S}, E_{\mathcal{S}})$ if $\{u, v\} \in \mathcal{S}$ and $u$ and $v$ connect to a same node $w \in \mathcal{D}$. Similarly, one could obtain one-mode projection graphs $G_{\mathcal{D}} = (\mathcal{D}, E_{\mathcal{D}})$ if $\{x, y\} \in \mathcal{D}$ and $x$ and $y$ connect to a same node $z \in \mathcal{S}$. One-mode projection graphs are widely used to capture the connections or associations among nodes in the same side of the bipartite graphs [13]. Thus, we refer to one-mode projection graphs $G_{\mathcal{S}} = (\mathcal{S}, E_{\mathcal{S}})$ and $G_{\mathcal{D}} = (\mathcal{D}, E_{\mathcal{D}})$ as *source behavior graph (SBG)* and *destination behavior graph (DBG)*, respectively. In this study, we leverage one-mode projection graphs to explore the social-behavior similarity among source or destination IP addresses associated with the same Internet applications.

Figure 1[a] illustrates an example of a simple bipartite graph that shows data communications between six source IP addresses and four destination IP addresses on a particular destination port. Figure 1[b] is the derived one-mode projection of the bipartite graph of Figure 1[a] on the six source addresses, while Figure 1[c] is a one-mode projection on the four destination addresses. Figures 1[b][c] reveal interesting observations in each side of the bipartite graphs, e.g., a clique in four source IP addresses ($s_1$, $s_2$, $s_5$, and $s_6$) in Figure 1[b]. Such observations could be used to discover interesting traffic patterns of Internet applications.




(a) An example of bipartite graphs (b) One-mode projection graph of source hosts (c) One-mode projection graph of destination hosts

Fig. 1. Modeling application traffic using bipartite graphs and one-mode projection graphs

### B. Coefficient Clustering of Behavior Graphs

Coefficient clustering is a widely-used measure to study the "closeness", or the "small-world" patterns of nodes in one-mode project graphs [14]. This measure can be applied to individual nodes as *local coefficient clustering (LCC)* and can also be applied to the entire graph as *global coefficient clustering (GCC)*. For a given node $u$, the local coefficient clustering, $LCC_u$, is provided by the number of the edges among $u$'s neighbors over the number of all possible edges among $u$'s neighbors. Let $N_u$ represent the set of all the neighbors of the node $u$, where $|N_u| = m$, and let $E_u$ represent the set of edges among these neighbors. The number of all possible edges among $m$ neighbors is $m \times (m-1)/2$. The *local coefficient clustering (LCC)* of $u$ is calculated as:

$$LCC_u = \frac{|E_u|}{(m*(m-1))/2} = \frac{|E_u|*2}{m*(m-1)}. \qquad (1)$$

Clearly $LCC_u \in [0,1]$. $LCC_u$ is 0 if there is no edges among $u$'s neighbors, while $LCC_u$ is 1 if $u$'s neighbors form a complete graph (clique). Note that the *local coefficient clustering (LCC)* for nodes with 0 or 1 neighbor is 0 due to zero edges. The *global coefficient clustering (GCC)* of the entire graph, $GCC_G$ is the average *local coefficient clustering (LCC)* over all $n$ nodes, where $GCC_G = \frac{1}{N} * \sum LCC_u$, where $u \in G$. Because of the existence of nodes with 0 or 1 neighbor which affect the calculation of the global coefficient clustering, we adopt an adaptive global coefficient clustering (AGCC), introduced in [15], $AGCC_G = \frac{1}{1-\theta} * GCC_G$, where $\theta$ is the percentage of the isolated nodes in one-mode project graphs. In addition, we also measure the percentage of the nodes that have at least two neighbors (or non-isolated nodes) in the graphs. Our experiment results show that coefficient clustering captures social behavior of source or destination IP addresses engaging in the same applications, thus are very useful for uncovering Internet applications exhibiting similar traffic patterns.

### C. Clustering Analysis of Internet Application

For the source and destination behavior graphs generated from the traffic of each Internet application, we calculate the *adaptive global coefficient clustering*. In the rest of this paper, we refer to the *adaptive global coefficient clustering* as *coefficient clustering* for simplicity.

Figure 2 shows the distribution of coefficient clustering for all Internet applications observed from an OC-192 Internet backbone link during a one-minute time window. An interesting observation is that the clustered pattern of coefficient clustering, which leads to our next step of applying clustering algorithms to discover the inherent clusters formed by Internet applications sharing similar behavior patterns.

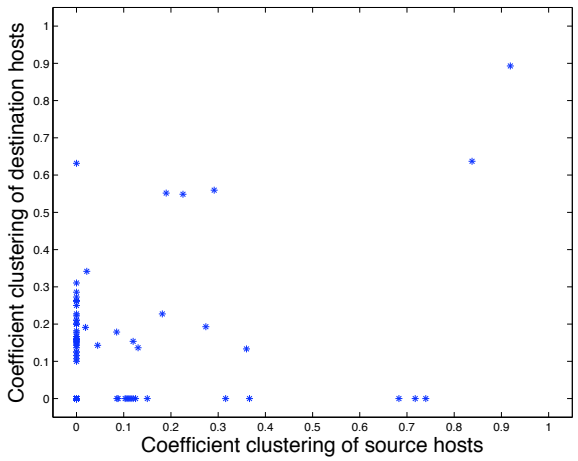



Fig. 2. Distribution of adaptive global coefficient clustering for source and destination behavior graphs of Internet applications

In this study, we apply a simple $K$-means clustering algorithm [16] on the feature space of Internet applications and to group them into distinct clusters with similar coefficient clustering. The choice of selecting this algorithm is due to its simplicity and wide usage. The features used in the clustering algorithm include coefficient clustering of source

and destination behavior graphs and the ratios of nodes with two or more neighbors in these graphs. In other words, for each source or destination port $p$, we obtain a vector of four features, i.e., $AGCC_{\mathcal{S}_p}$, $AGCC_{\mathcal{D}_p}$, $r_{\mathcal{S}_p}$ $r_{\mathcal{D}_p}$), where the first two features are coefficient clustering of source and destination hosts engaging in the application port $p$ and the last two features are ratios of hosts with at least two or more neighbors in one-mode project graphs on source and destination hosts.

A challenging issue of applying $K$-means clustering algorithms is to find an optimal value of $k$, since the choice of $k$ plays an important role of archiving the high quality of clustering results. In this study, we search the optimal value of $k$ by running $K$-means algorithms using a variety of $k$ values and evaluate the best choice of $k$ by comparing the sum of squared error (SSE) with Euclidian distance function between nodes in each cluster [16]. For example, Figure 3 illustrates the distribution of SSE with varying values of $k$ from 1 to 16. We select $k = 9$ as the choice since increasing $k$ from 9 to 10 and above does not bring significant benefits of reducing SSE. In the next section, we will present experiments results of applying the proposed method on real network traffic collected from Internet backbone links in a large ISP network.

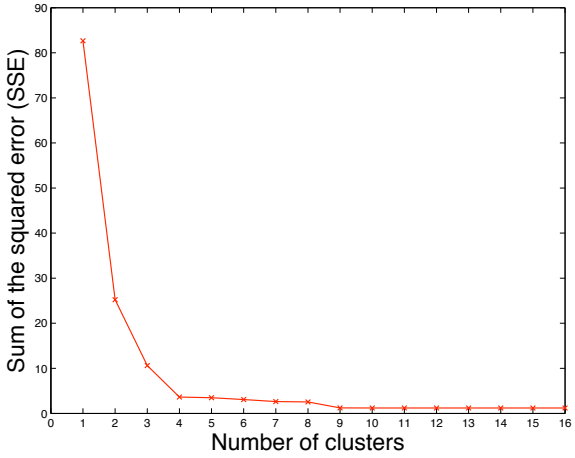



Fig. 3. Determining the optimal $k$ based on sum of squared error (SSE)

## IV. Experiment Results

In this section, we first present the general observations of Internet applications observed from Internet backbone links, and then we discuss distinctive traffic characteristics of applications clusters. We conclude this section by demonstrating the usage of behavioral graph analysis for detecting emerging applications and anomalous traffic patterns towards Internet applications .

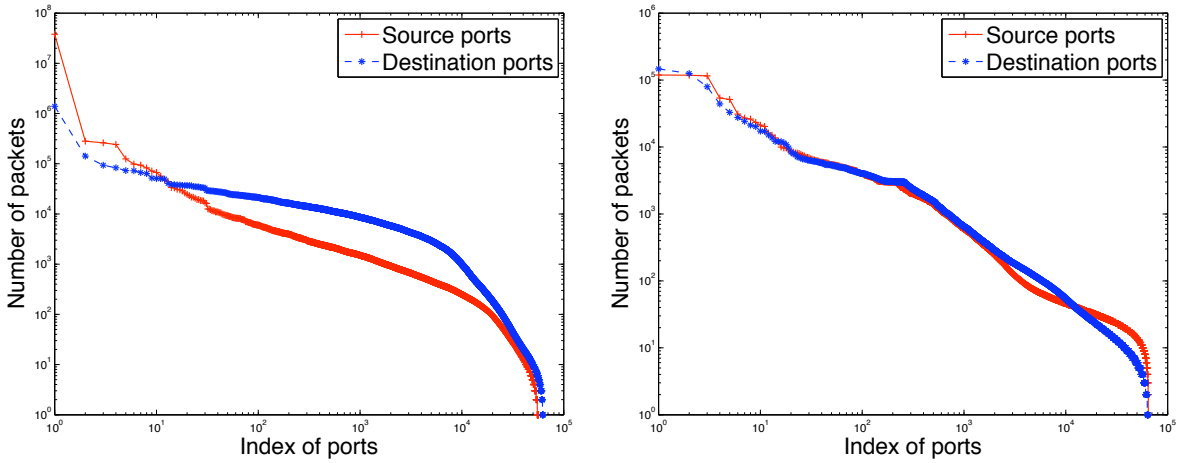
### A. Traffic Characteristics of Internet Applications

The network traffic datasets used in this study are collected from bidirectional OC-192 Internet backbone links in a large Internet service provider through CAIDA's equinix-chicago and equinix-sanjose network monitors [17] on December 17, 2009. Due to privacy reasons, the traffic traces are anonymized using CryptoPAn *prefix-preserving* anonymization method [18]. The *prefix-preserving* process does not affect the analysis in this study, since our analysis is focused on the social behaviors of all source or destination hosts engaging in the same Internet applications.

Figures 4[a][b] illustrate the distribution of IP packets for Internet application traffic observed from one backbone link during one-minute time window for TCP and UDP ports, respectively. It is interesting to observe that a large number of application ports, regardless transport protocols (TCP or UDP) and traffic directions (source ports or destination ports), carry non-trivial data traffic. For example, there are over 2550 TCP destination ports with more than 5000 IP packets on the link during the one minute time window. In other words, the traditional top $N$ approaches of focusing on a few top ports with the largest amount of traffic is not sufficient, since it is also very important to study the other applications with significant volumes of IP traffic. Another interesting observation we find is the wide diversity of applications among the TCP or UDP ports. For example, Table I lists the top 10 ports/protocol for source and destination ports. Besides several major applications such as Web, DNS, SMTP, P2P, many of these top ports are associated with unknown applications. Thus, the diverse applications and vast amount traffic highlight the importance of examining a broad range of Internet applications, and call for new scalable methods to study traffic behaviors of Internet applications

Building source and destination behavior graphs for each application port in our proposed method provides an opportunity to understand the social behavior of source and destination hosts engaging in the same applications. In addition, grouping these applications based on coefficient clustering of source and destination behavior graphs into distinct clusters helps understand unknown applications that share similar patterns with well-known applications. To evaluate the quality of the clustering results, we study traffic characteristics of application clusters and compare the similarity in traffic characteristics among application ports in the same clusters as well as the dissimilarity among ports in different clusters.

Our preliminary results show that the application clusters indeed exhibit distinctive traffic characteristics. Specifically, for each application port $p$, we study IP symmetry $ipsym_p$, fan-out degree of source hosts $fanout_p$ and fan-in degree of destination hosts $fanin_p$. The IP symmetry $ipsym_p$ is given by the ratio between unique source hosts and unique destination hosts engaging in the application port $p$. The fan-out degree for the application is the average fan-out degree of all source hosts involving in the application, while the fan-in degree is the average fan-in degree of all destination hosts. Figures 5[a-c] illustrate the distinctive characteristics in IP symmetry, fan-out degrees of source hosts, and fan-in degree of destination hosts for application clusters during one time window, respectively. The similar observations hold for other time windows as well. This observation confirms that our proposed method of behavioral graph analysis on Internet applications indeed is able to discover distinct clusters




(a) Application ports with TCP protocol

(b) Application ports with UDP protocol

Fig. 4. Distribution of network traffic for Internet applications observed from Internet backbone links

TABLE I
TOP 10 PORTS/PROTOCOLS OF INTERNET APPLICATION TRAFFIC BASED ON OBSERVED IP PACKETS FROM AN INTERNET BACKBONE LINK DURING ONE-MINUTE TIME WINDOW.

| Top N | Source Ports | | Destination Ports | |
|---|---|---|---|---|
| | TCP | UDP | TCP | UDP |
| 1 | 80 (HTTP) | 12340 (Skype) | 80 (HTTP) | 3074 (Xbox LIVE) |
| 2 | 443 (HTTPS) | 53 (DNS) | 6346 (GNUtella) | 53 (DNS) |
| 3 | 25 (SMTP) | 3074 (Xbox LIVE) | 51413 (BitTorrent) | 6346 (GNUtella) |
| 4 | 1935 (Streaming) | 80 (Unknown) | 25 (SMTP) | 15000 (Unknown) |
| 5 | 60020 (Unknown) | 15000 (Unknown) | 445 (Windows SMB) | 80 (Unknown) |
| 6 | 554 (Streaming) | 6881 (BitTorrent) | 6699 (WinMX) | 6257 (WinMX) |
| 7 | 9050 (Unknown) | 13001 (Unknown) | 443 (HTTPS) | 1484 (Unknown) |
| 8 | 27030 (Gaming) | 6257 (WinMX) | 50159 (Unknown) | 3478 (Unknown) |
| 9 | 6699 (WinMX) | 54090 (Unknown) | 49634 (Unknown) | 12340 (Skype) |
| 10 | 800 (Unknown) | 12035 (Gaming) | 50514 (Unknown) | 3658 (Unknown) |

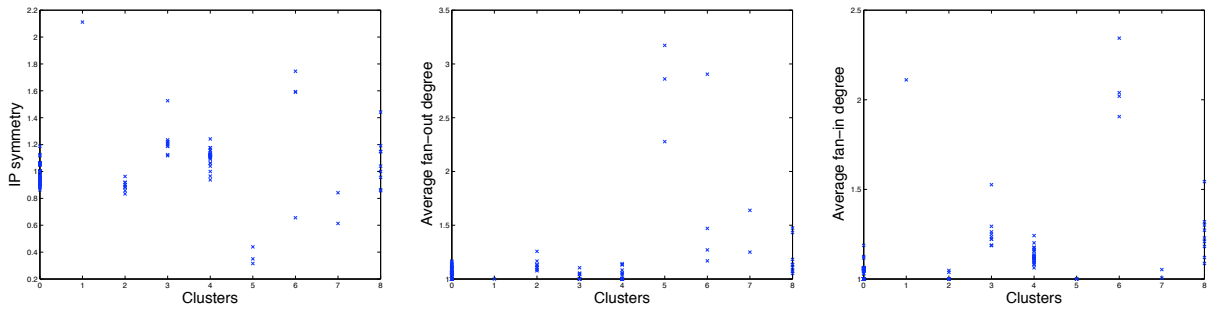
of application ports that not only exhibit similar coefficient clustering in their source and destination behavior graphs, but also share similar traffic characteristics in IP symmetry, fan-out and fan-in degrees.

### *B. Applications*

To demonstrate the usage of behavioral graph analysis, we use case studies to illustrate how the proposed method could aid in identifying emerging applications and detecting anomalous traffic patterns.

**Detecting Emerging Applications:** As the Internet continues to grow in end users, mobile devices, and applications, classifying Internet applications becomes more complicated due to the rapid growth of new applications, mixed uses of application ports, traffic hiding using well-known ports for avoiding firewall filtering. On the other hand, detecting emerging applications is very important for traffic engineering and security monitoring. In our preliminary analysis, we find that application clusters are a feasible approach of finding new applications that share similar coefficient clustering or similar social behaviors of end hosts with existing known applications. For example, we find that an unknown port consistently follows in the same clusters with service ports TCP port 25 (SMTP), TCP port 80 (Web), TCP port 443 (HTTPS). We conjecture that this port very likely corresponds to a service application since the source and destination hosts engaging in this port exhibit similar social behaviors with existing known applications. Such findings could provide very valuable information for network operators for in-depth analysis.

**Detecting Anomalous Traffic Patterns:** The applications of behavior graphs analysis on Internet application traffic also include detecting anomalous traffic patterns. For example, the clustering results of application traffic shows a cluster of TCP destination ports 135, 1433, and 22. The first two are ports associated with well-known vulnerabilities, thus it is not surprising to observe these two ports in the same cluster. However, port 22 is mostly used for SSH traffic, and it is expected to be grouped into clusters that include other major Internet service ports. An in-depth analysis reveal that during that particular time window, six source IP addresses in the same `/26` network prefix send only TCP SYN packet to 28 unique destination address on destination TCP port 22. These hosts likely scan SSH ports on Internet hosts. The scanning traffic together account for 66% of total flows towards destination TCP port 22 during the time window,




(a) IP symmetry (b) Fan-out degree of source IP addresses (c) Fan-in degree of destination IP addresses

Fig. 5. Distinctive traffic characteristics of application clusters

which explains why TCP port 22 is clustered with ports associated with well-known vulnerabilities.

In summary, our experiment results show the proposed method of behavioral graph analysis is able to group Internet ports into distinct clusters based on coefficient clustering and other graph properties of source and destination behavior graphs. The application clusters could aid network operators in understanding emerging applications and detecting anomalous traffic towards Internet applications.

## V. Conclusions and Future Work

This paper uses bipartite graphs and one-mode projection graphs to analyze social behavior of end hosts engaging in the same Internet applications. Through coefficient clustering and other graph properties, we find interesting similarity of social behavior among different applications, and then apply a simple $K$-means clustering algorithm to group applications with similar social behavior into distinctive clusters. Further in-depth analysis confirms that Internet applications in the same clusters also have similar characteristics in the aggregated traffic. In addition, we demonstrate the usage of the proposed method of behavioral graph analysis in detecting emerging applications and anomalous traffic behaviors towards Internet applications. We are currently in the process of designing and implementing a prototype system to evaluate real-time performance and cost of the proposed method. In addition, we are also interested in analyzing and correlating social behavior of Internet applications from multiple backbone links.

## References

[1] C. Shue, A. Kalafut, and M. Gupta, "The web is smaller than it seems," in *Proceedings of ACM SIGCOMM conference on Internet measurement*, October 2007.

[2] D. Schatzmann, W. Muehlbauer, T. Spyropoulos, and X. Dimitropoulos, "Digging into HTTPS: Flow-Based Classification of Webmail Traffic," in *Proceedings of ACM SIGCOMM conference on Internet measurement*, November 2010.

[3] D. Plonka and P. Barford, "Context-aware Clustering of DNS Query Traffic," in *Proceedings of ACM Internet Measurement Conference*, October 2008.

[4] T.S. Choi, C.H. Kim, S. Yoon, J.S. Park, B.J. Lee, H.H. Kim, H.S. Chung, and T.S. Jeong, "Content-aware Internet application traffic measurement and analysis," in *Proceedings of IEEE/IFIP Network Operations and Management Symposium (NOMS)*, April 2004.

[5] S. Wei, J. Mirkovic, and E. Kissel, "Profiling and Clustering Internet Hosts," in *Proceedings of the International Conference on Data Mining*, June 2006.

[6] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," in *Proceedings of ACM SIGCOMM*, August 2005.

[7] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Internet Traffic Behavior Profiling for Network Security Monitoring," *IEEE/ACM Transactions on Networking*, vol. 16, no. 6, pp. 1241–1252, December 2008.

[8] H. Jiang, Z. Ge, S. Jin, and J. Wang, "Network Prefix-level Traffic Profiling: Characterizing, Modeling, and Evaluation," *Computer Networks*, 2010.

[9] M. Schwartz, and D. Wood, "Discovering shared interests using graph analysis," *Communications of the ACM*, vol. 36, no. 8, pp. 78 – 89, August 1993.

[10] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, and G. Varghese, "Network monitoring using traffic dispersion graphs," in *Proceedings of ACM SIGCOMM Internet Measurement Conference*, 2007.

[11] Y. Jin, E. Sharafuddin, and Z.-L. Zhang, "Unveiling core network-wide communication patterns through application traffic activity graph decomposition," in *Proceedings of ACM SIGMETRICS*, June 2009.

[12] K. Xu, F. Wang, and L. Gu, "Network-Aware Behavior Clustering of Internet End Hosts," in *Proceedings of IEEE INFOCOM*, April 2011.

[13] J.-L. Guillaume, and M. Latapy, "Bipartite graphs as models of complex networks," *Physica A: Statistical and Theoretical Physics*, vol. 371, no. 2, pp. 795 – 813, 2006.

[14] J. Ramasco, S. Dorogovtsev, and P. Romualdo, "Self-organization of collaboration networks," *Physical Review*, vol. 70, no. 3, 2004.

[15] M. Kaiser, "Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks," *New Journal of Physics*, vol. 10, August 2008.

[16] P.-N. Tan, M. Steinbach, and Vipin Kumar, *Introduction to Data Mining*. Addison-Wesley, 2006.

[17] Cooperative Association for Internet Data Analysis (CAIDA), "Internet Traces," http://www.caida.org/data/passive/passive_2009_dataset.xml.

[18] J. Fan, J. Xu, M. Ammar, and S. Moon, "Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme," *Computer Networks*, vol. 46, no. 2, pp. 253–272, October 2004.